CS6350 – Big Data Management and Analysis

Assignment submitted by Yash Vijaynarayan Gupta

WordCount for Named Entities

Problem Statement - In this part, you will compute the word frequency for named entities in a large file. You are free to use any NLP library that works with Scala/PySpark. Some examples are:

- NLTK library https://www.nltk.org/
- John Snow Labs https://github.com/JohnSnowLabs/spark-nlp-workshop

Solution:

The assignment is done using Databricks platform.

Steps of the assignment would be as follows:

1. Find a large text file from the Gutenberg project: https://www.gutenberg.org and upload it to your Databricks cluster. In our case (sherlock.txt)
2. Use an NLP library (NLTK) to extract only the named entities from the text.
3. Write code for a map-reduce program that performs wordcount on the extracted named entities.
4. The output from the map task should be in the form of (key, Value) where key is the named-entity, and value is its count (i.e. once every time it occurs)
5. The output from the reducer should be sorted in descending order of count. That is, the named-entity that is most frequent should appear at the top.